COMBINATORIAL PROBLEMS RELATED TO GEOMETRICALLY DISTRIBUTED RANDOM VARIABLES

Helmut Prodinger

Department of Algebra and Discrete Mathematics
Technical University of Vienna, Austria

1. Introduction

Motivated from Computer Science problems we consider the following situation (compare [2] and [3]). In these papers the reader will find a more complete description as well as additional references.

Let X denote a geometrically distributed random variable, i.e. $\mathbb{P}\{X = k\} = pq^{k-1}$ for $k \in \mathbb{N}$ and $q = 1 - p$. Assume that we have n independent random variables $X_1, \dots, X_n$ according to this distribution.

The first parameter of interest is the *number of left-to-right maxima*, where we say that X_i is a left-to-right maximum (in the strict sense) if it is strictly larger than the elements to left. A left-to-right maximum in the loose sense is defined analogously but "larger" is replaced by "larger or equal".

The second parameter of interest is the *(horizontal) path length*, i.e. the sum of the left-to-right maxima in the loose sense of all the sequences $X_i, \dots, X_n$, where i is running from 1 to n.

Example. Consider the sequence $4, 5, 2, 3, 5$. It has 2 left-to-right maxima in the strict sense (4–5) and 3 left-to-right maxima in the loose sense (4–5–5). For the path length we must consider the subsequences

$$\begin{aligned}
&4, 5, 2, 3, 5\\
&5, 2, 3, 5\\
&2, 3, 5\\
&3, 5\\
&5
\end{aligned}$$

with respectively $3, 2, 3, 2, 1$ left-to-right maxima. Therefore the path length is $3 + 2 + 3 + 2 + 1 = 11$.

2. Left-to-right maxima in the strict sense

We can find a probability generating function by considering an appropriate "language".

The "letters" will be denoted by $\mathbf{1}, \mathbf{2}, \ldots$. We decompose all sequences $X_1 X_2 \ldots$ in a canonical way as follows: We combine each left-to-right maximum $\mathbf{k}$ with the following (smaller or equal) elements. Such a part is decribed by

$$\mathcal{A}_k := \mathbf{k}\{\mathbf{1}, \ldots, \mathbf{k}\}^*.$$

Such a group may be present or not. This observation gives the desired "language", where ε denotes the "empty word":

$$\mathcal{L} := \big(\mathcal{A}_1 + \varepsilon\big) \cdot \big(\mathcal{A}_2 + \varepsilon\big) \cdot \big(\mathcal{A}_3 + \varepsilon\big) \ldots$$

Now we want to mark each letter by a "z" and each left-to-right maximum by a "y". The probability pq^{k-1} for a letter $\mathbf{k}$ should of course not being forgotten. $\{\mathbf{1}, \ldots, \mathbf{k}\}$ maps into $z(1-q^k)$ and its star $\{\mathbf{1}, \ldots, \mathbf{k}\}^*$ into $1/\big(1 - z(1-q^k)\big)$. So we obtain the generating function $F(z, y)$ as an infinite product:

$$F(z, y) = \prod_{k \geq 1} \left(1 + \frac{yzpq^{k-1}}{1 - z(1-q^k)}\right)$$

To be explicit, the coefficient of $z^n y^k$ in $F(z, y)$ is the probability that n random variables have k left-to-right maxima.

Observe that, as it is to be expected, $F(z, 1) = \dfrac{1}{1-z}$, as it is then a telescoping product.

Let $f(z) = \frac{\partial F(z,y)}{\partial y}\Big|_{y=1}$. It is the generating function for the expected values E_n, i.e. the $E_n = [z^n]f(z)$. Performing this differentiation we are led to

$$f(z) = \frac{pz}{1-z} \sum_{k \geq 0} \frac{q^k}{1 - z(1-q^k)},$$

which is also, by partial fraction decomposition,

$$f(z) = p \sum_{k \geq 0} \left[\frac{1}{1-z} - \frac{1}{1 - z(1-q^k)}\right].$$

From this the coefficients E_n are easy to see, because there are only geometric series:

$$E_n = [z^n]f(z) = p \sum_{k \geq 0} \Big[1 - \big(1-q^k\big)^n\Big] = p \sum_{k=1}^{n} \binom{n}{k} (-1)^{k-1} \frac{1}{1-q^k}$$

The asymptotic evaluation of such an alternating sum is conveniently performed by *Rice's method*, which we cite as a lemma.

Lemma. *Let $\mathcal{C}$ be a curve surrounding the points $1, 2, \ldots, n$ in the complex plane and let $f(z)$ be analytic inside $\mathcal{C}$. Then*

$$\sum_{k=1}^{n} \binom{n}{k} (-1)^k f(k) = -\frac{1}{2\pi i} \int_{\mathcal{C}} [n; z] f(z) dz,$$

where

$$[n;z] = \frac{(-1)^{n-1}n!}{z(z-1)\dots(z-n)} = \frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)} = B(n+1,-z). \quad \square$$

Extending the contour of integration it turns out that under suitable growth conditions on $f(z)$ the asymptotic expansion of the alternating sum is given by

$$\sum \operatorname{Res}\big([n;z]f(z)\big) + \text{smaller order terms}$$

where the sum is taken over all poles z_0 different from $1,\dots,n$.

The range $1,\dots,n$ for the summation is not sacred; if we sum, for example, over $k=2,\dots,n$, the contour must encircle $2,\dots,n$, etc.

Theorem 1. *The average number E_n of left-to-right maxima (strict model) in the context of n independently distributed geometric random variables has the asymptotic expansion*

$$E_n = p\left[\log_Q n + \frac{\gamma}{L} + \frac{1}{2} - \delta\left(\log_Q n\right)\right] + O\big(\frac{1}{n}\big)$$

where $Q=q^{-1}$, $L=\log Q$, γ is Euler's constant and δ is a periodic function of period 1, mean zero and small amplitude. Its Fourier series is given by

$$\delta(x) = \frac{1}{L}\sum_{k\neq 0}\Gamma\left(-\chi_k\right)e^{2k\pi i x}.$$

The variance can also be computed, by considering the second derivative of $F(z,y)$ with respect to y.

Theorem 2. *The variance V_n of the number of left-to-right maxima (strict model) in the context of n independently distributed geometric random variables has the asymptotic expansion for $n\to\infty$*

$$V_n = pq\log_Q n + p^2\left(-\frac{5}{12} + \frac{\pi^2}{6L^2} - \frac{\gamma}{L} - \left[\delta^2\right]_0\right) + p\left(\frac{\gamma}{L}+\frac{1}{2}\right) + \delta_1(\log_Q n) + O\big(\frac{1}{n}\big).$$

Here, $\left[\delta^2\right]_0$ is the mean of the square of $\delta^2(x)$, a very small quantity that can be neglected for numerical purposes. Furthermore, $\delta_1(x)$ is a periodic function with mean 0; its Fourier coefficients could be described if needed.

3. Left-to-right maxima in the loose sense

Again, we are defining an appropriate "language" $\mathcal{L}$ from which a bivariate generating function $F(z,y)$ can be derived.

Set $\mathcal{A}_k := \mathbf{k}\{\mathbf{1},\dots,\mathbf{k-1}\}^*$, then $\mathcal{L} := \mathcal{A}_1^*\cdot\mathcal{A}_2^*\cdot\mathcal{A}_3^*\dots$ and

$$F(z,y) = \prod_{k\geq 1}\frac{1}{1-\dfrac{yzpq^{k-1}}{1-z(1-q^{k-1})}} = \prod_{k\geq 0}\frac{1-z(1-q^k)}{1-z+zq^k(1-py)}.$$

Therefore

$$f(z) = \frac{\partial F(z,y)}{\partial y}\bigg|_{y=1} = \frac{pz}{1-z}\sum_{k\geq 0}\frac{q^k}{1-z(1-q^{k+1})} = \frac{p}{q}\sum_{k\geq 1}\left[\frac{1}{1-z} - \frac{1}{1-z(1-q^k)}\right]$$

and

$$E_n = [z^n]f(z) = \frac{p}{q}\sum_{k\geq 1}\left[1-\left(1-q^k\right)^n\right].$$

Theorem 3. *The average number E_n of left-to-right maxima (loose model) in the context of n independently distributed geometric random variables has the asymptotic expansion*

$$E_n = \frac{p}{q}\left[\log_Q n + \frac{\gamma}{L} - \frac{1}{2} - \delta\left(\log_Q n\right)\right] + O\big(\frac{1}{n}\big).$$

Theorem 4. *The variance V_n of the number of left-to-right maxima (loose model) in the context of n independently distributed geometric random variables has the asymptotic expansion for $n \to \infty$*

$$V_n = \frac{p}{q^2}\log_Q n + \frac{p^2}{q^2}\left(-\frac{5}{12} + \frac{\pi^2}{6L^2} + \frac{\gamma}{L} - \frac{2}{L} - \left[\delta^2\right]_0\right) + \frac{p}{q}\left(\frac{\gamma}{L} - \frac{1}{2}\right) + \delta_2(\log_Q n) + O\big(\frac{1}{n}\big).$$

Here, $\left[\delta^2\right]_0$ is the mean of the square of $\delta^2(x)$, a very small quantity that can be neglected for numerical purposes. Furthermore, $\delta_2(x)$ is a periodic function with mean 0; its Fourier coefficients could be described if needed.

4. Path length

If we denote the path length of a "word" ω by $a(\omega)$, then we have the following recursion formula

$$a(\omega) = a(\rho m \sigma) = a(\rho) + a(\sigma) + 1 + |\rho|$$

with $\rho \in \{\mathbf{1}, \dots, \mathbf{m}\}^*$ and $\sigma \in \{\mathbf{1}, \dots, \mathbf{m} - 1\}^*$. From this we get a functional equation for the generating functions. (The upper index '$= m$' e.g. refers to all sequences where the maximal element is m.)

$$P^{=m}(z,y) = pq^{m-1} zy P^{\le m}(zy, y) P^{<m}(z,y)$$

It is not likely that this formidable equation can be solved explicitly. However it contains enough information to obtain the generating functions for the expectations (and the variance, too). For the expectation we have to differentiate with respect to y and then set $y = 1$. Denoting the corresponding functions by $F^*(z)$, we get

$$\begin{aligned}\frac{F^{=m}(z)}{pq^{m-1}z} &= P^{\le m}(z,1)P^{<m}(z,1) \\ &\quad + \left[z\frac{\partial}{\partial z}P^{\le m}(z,1) + F^{\le m}(z)\right]P^{<m}(z,1) + P^{\le m}(z,1)F^{<m}(z).\end{aligned}$$

Since $\quad P^{\le m}(z,1) = \dfrac{1}{1 - z(1-q^m)} =: \dfrac{1}{[\![m]\!]} \quad$ and $\quad \dfrac{\partial}{\partial z}P^{\le m}(z,1) = \dfrac{1-q^m}{[\![m]\!]^2}$,
we obtain

$$F^{\le m}(z)[\![m]\!]^2 = F^{<m}(z)[\![m-1]\!]^2 + \frac{pq^{m-1}z}{[\![m]\!]},$$

which we can solve by iteration:

$$F^{\le m}(z) = \frac{p}{q}\frac{z}{[\![m]\!]^2}\sum_{i=1}^{m}\frac{q^i}{[\![i]\!]}$$

The limit for $m \to \infty$ is the generating function for the expectations:

$$F(z) = \frac{p}{q} \frac{z}{(1-z)^2} \sum_{i \geq 1} \frac{q^i}{[\![i]\!]}$$

From this it is easy to get the coefficients E_n

$$E_n = [z^n]F(z) = \frac{p}{q} \sum_{k=2}^{n+1} \binom{n+1}{k} (-1)^k \frac{1}{Q^{k-1} - 1}$$

Theorem 5. *The expected path length E_n in the context of n independently distributed geometric random variables has the asymptotic expansion*

$$E_n = (Q-1)n\Big(\log_Q n + \frac{\gamma - 1}{L} - \frac{1}{2} + \frac{1}{L}\delta_3(\log_Q n)\Big) + O(\log n)$$

with $\quad \delta_3(x) = \sum_{k \neq 0} \Gamma(-1 - \frac{2k\pi i}{L}) e^{2k\pi i x}.$

To deal with the variance, we have to differentiate the functional equation twice. Denoting the resulting generating function by $H(z)$, we finally find

$$\begin{aligned}
H(z) &= 2\frac{p}{q}\frac{z}{(1-z)^2} \sum_{i \geq 1} \frac{q^i}{[\![i]\!]^2} \\
&- 2\frac{p}{q}\frac{z}{(1-z)^2} \sum_{i \geq 1} \frac{q^i}{[\![i]\!]} \\
&- 2\Big(\frac{p}{q}\Big)^2 \frac{z^2}{(1-z)^2} \sum_{1 \leq j \leq i} \frac{q^{i+j}}{[\![i]\!][\![j]\!]} \\
&+ 4\Big(\frac{p}{q}\Big)^2 \frac{z^2}{(1-z)^2} \sum_{1 \leq j \leq i} \frac{q^{i+j}}{[\![i]\!]^2[\![j]\!]} \\
&+ 2\Big(\frac{p}{q}\Big)^2 \frac{z^2}{(1-z)^2} \sum_{1 \leq j \leq i} \frac{q^{i+j}}{[\![i]\!][\![j]\!]^2} \\
&+ 2\Big(\frac{p}{q}\Big)^2 \frac{z^2}{(1-z)^2} \sum_{1 \leq j < i} \frac{q^{i+j}}{[\![i]\!][\![i-1]\!][\![j]\!]} \\
&+ 2\Big(\frac{p}{q}\Big)^3 \frac{z^3}{(1-z)^2} \sum_{i \geq 1} \sum_{1 \leq j \leq i} \sum_{1 \leq h < i} \frac{q^{i+j+h}}{[\![i]\!][\![i-1]\!][\![j]\!][\![h]\!]}
\end{aligned}$$

To get the coefficient of z^n in an efficient way we use to following principle (used recently by Flajolet and Richmond).

$$A(z) = \sum_{n \geq 0} a_n z^n \longleftrightarrow \frac{1}{1-w} A\Big(\frac{w}{1-w}\Big) = \sum_{n \geq 0} \Big(\sum_{k=0}^{n} \binom{n}{k} a_k\Big) w^n$$

Of course this relation can be inverted to read

$$A(z) = \sum_{n \geq 0} \Big(\sum_{k=0}^{n} \binom{n}{k} (-1)^k f(k)\Big) z^n \longleftrightarrow \frac{1}{1-w} A\Big(\frac{w}{w-1}\Big) = \sum_{n \geq 0} f(n) w^n.$$

That means that if we find the coefficients in the "w–world" we *automatically* have the coefficients in the "z–world" as *alternating sums*!

This is especially convenient, since the expressions become nicer if we substitute $z = \dfrac{w}{w-1}$, because $\dfrac{1}{[\![i]\!]} = \dfrac{1-w}{1-wq^i}$.

Consider as an example the first one, namely $\quad A(z) = \dfrac{z}{(1-z)^2} \sum_{i\ge1} \dfrac{q^i}{[\![i]\!]^2}$.

We easily find that

$$\frac{1}{1-w} A\Big(\frac{w}{w-1}\Big) = -(1-w)^2 \sum_{i\ge1} \frac{wq^i}{\big(1-wq^i\big)^2}.$$

Let us forget the extra factor $-(1-w)^2$ for a moment. We find for $n \ge 1$

$$[w^n] \sum_{i\ge1} \frac{wq^i}{\big(1-wq^i\big)^2} = \sum_{i\ge1} q^{in} \cdot [w^n] \frac{w}{(1-w)^2} = \frac{n}{Q^n-1}.$$

But because $(1-w)^2 = \dfrac{w^2}{z^2}$, the extra factor just works as a *shift* on both sides, so that we find

Sum 1.

$$[z^n] \frac{z}{(1-z)^2} \sum_{i\ge1} \frac{q^i}{[\![i]\!]^2} = -\sum_{k=3}^{n+2} \binom{n+2}{k} (-1)^k \frac{k-2}{Q^{k-2}-1}$$

In this way we find for the coefficients of z^n in the other sums in $H(z)$ the following expressions as alternating sums.

Sum 2.

$$[z^n] \frac{z}{(1-z)^2} \sum_{i\ge1} \frac{q^i}{[\![i]\!]} = \sum_{k=2}^{n+1} \binom{n+1}{k} (-1)^k \frac{1}{Q^{k-1}-1}$$

Sum 3.

$$[z^n] \frac{z^2}{(1-z)^2} \sum_{1\le j\le i} \frac{q^{i+j}}{[\![i]\!][\![j]\!]} = -\sum_{k=3}^{n+1} \binom{n+1}{k} (-1)^k \frac{1}{Q^{k-1}-1} \left[k-2 + \sum_{m=1}^{k-2} \frac{1}{Q^m-1} \right]$$

Sum 4.

$$[z^n] \frac{z^2}{(1-z)^2} \sum_{1\le j\le i} \frac{q^{i+j}}{[\![i]\!]^2[\![j]\!]} = \sum_{k=4}^{n+2} \binom{n+2}{k} (-1)^k \frac{1}{Q^{k-2}-1} \left[\binom{k-2}{2} + \sum_{m=1}^{k-3} \frac{m}{Q^m-1} \right]$$

Sum 5.

$$[z^n] \frac{z^2}{(1-z)^2} \sum_{1\le j\le i} \frac{q^{i+j}}{[\![i]\!][\![j]\!]^2}$$

$$= \sum_{k=4}^{n+2} \binom{n+2}{k} (-1)^k \frac{1}{Q^{k-2}-1} \left[\binom{k-2}{2} + (k-2) \sum_{m=1}^{k-3} \frac{1}{Q^m-1} - \sum_{m=1}^{k-3} \frac{m}{Q^m-1} \right]$$

Sum 6.

$$[z^n]\frac{z^2}{(1-z)^2}\sum_{1\le j<i}\frac{q^{i+j}}{[\![i]\!][\![i-1]\!][\![j]\!]}=\sum_{k=4}^{n+2}\binom{n+2}{k}(-1)^k\frac{k-3}{(Q-1)\left(Q^{k-2}-1\right)}$$

Sum 7.

$$[z^n]\frac{z^3}{(1-z)^2}\sum_{\substack{i\ge1\\1\le j\le i\\1\le h<i}}\frac{q^{i+j+h}}{[\![i]\!][\![i-1]\!][\![j]\!][\![h]\!]}$$

$$=-\sum_{k=5}^{n+2}\binom{n+2}{k}(-1)^k\frac{1}{(Q-1)\left(Q^{k-2}-1\right)}\left[\binom{k-3}{2}+\frac{k-4}{Q-1}+\sum_{m=2}^{k-3}\frac{m-2}{Q^m-1}\right]$$

This gives finally the following result:

Theorem 6. *The variance V_n of the path length in the context of n independently distributed geometric random variables has the asymptotic expansion*

$$V_n=(Q-1)^2n^2\left\{\frac{Q+1}{2(Q-1)L}+\frac{1}{L^2}-\frac{\pi^2}{6L^2}+\frac{8\pi^2}{L^2}h\Big(\frac{4\pi^2}{L}\Big)-\alpha_1+\delta_4(\log_Q n)\right\}$$
$$+O\big(n^{1+\varepsilon}\big),\qquad \varepsilon>0,$$

where $h(x)=\sum_{k\ge1}\frac{e^{kx}}{\left(e^{kx}-1\right)^2}$, $\alpha_1=L\sum_{k\ge1}\frac{1}{k(L^2+4k^2\pi^2)\sinh(2k\pi^2/L)}$ *and $\delta_4(x)$ is again continuous, periodic of period 1 and mean zero.*

Both, $h(\cdot)$ and α_1 are very small quantities, so that a less accurate but more readible formula is

$$V_n\sim(Q-1)^2n^2\left\{\frac{Q+1}{2(Q-1)L}+\frac{1}{L^2}-\frac{\pi^2}{6L^2}\right\}.$$

5. A combinatorial interpretation of Euler's partition identities

The identities in question are (compare [1])

$$\prod_{n\ge0}\left(1+tq^n\right)=\sum_{n\ge0}\frac{t^nq^{\binom{n}{2}}}{Q_n(q)}$$

and

$$\prod_{n\ge0}\frac{1}{1-tq^n}=\sum_{n\ge0}\frac{t^n}{Q_n(q)},$$

where

$$Q_n(q)=\left(1-q^1\right)\left(1-q^2\right)\dots\left(1-q^n\right).$$

Now consider

$$\mathbb{P}\{X_1<\dots<X_n\}$$

and its generating function

$$M_<(z) = \sum_{n\ge0} \mathbb{P}\big\{X_1 < \cdots < X_n\big\} z^n$$

resp. the analogous quantities

$$\mathbb{P}\big\{X_1 \le \cdots \le X_n\big\}$$

and

$$M_\le(z) = \sum_{n\ge0} \mathbb{P}\big\{X_1 < \cdots < X_n\big\} z^n.$$

Then we can set up appropriate languages

$$\mathcal{M}_< = \big(\varepsilon + \mathbf{1}\big)\big(\varepsilon + \mathbf{2}\big)\dots$$

and

$$\mathcal{M}_\le = \mathbf{1}^* \cdot \mathbf{2}^* \dots,$$

so that

$$M_<(z) = \prod_{k\ge1} \big(1 + pq^{k-1}z\big)$$

resp.

$$M_\le(z) = \prod_{k\ge1} \frac{1}{1 - pq^{k-1}z}.$$

Using the identities we can further write

$$M_<(z) = \prod_{k\ge0} \big(1 + pq^k z\big) = \sum_{n\ge0} \frac{p^n z^n q^{\binom{n}{2}}}{Q_n(q)}$$

and

$$M_\le(z) = \prod_{k\ge0} \frac{1}{1 - pq^k z} = \sum_{n\ge0} \frac{p^n z^n}{Q_n(q)},$$

so that we have the *explicit* formulæ

$$\mathbb{P}\big\{X_1 < \cdots < X_n\big\} = \frac{p^n q^{\binom{n}{2}}}{Q_n(q)}$$

and

$$\mathbb{P}\big\{X_1 \le \cdots \le X_n\big\} = \frac{p^n}{Q_n(q)}.$$

References

1. G. Andrews, *The Theory of Partitions*, Addison-Wesley, Reading, Mass., 1976.
2. P. Kirschenhofer and H. Prodinger, *The path length of random skip lists*, Acta Informatica **31** (1994), 775–792.
3. H. Prodinger, *Combinatorics of geometrically distributed random variables: Left-to-right maxima*, 5th conference on Formal Power Series and Algebraic Combinatorics, Firenze (Discrete Mathematics, to appear) (1993).

TU Vienna
Wiedner Hauptstrasse 8–10
A-1040 Vienna
Austria

E-mail address: `proding@rsmb.tuwien.ac.at`